The THAI Journal of SURGERY

Official Publication of the Royal College of Surgeons of Thailand

Vol. 32 January - March 2011 No. 1

Review Article

Principles of Statistics for Surgeons II Descriptive or Summary Statistics

Panuwat Lertsithichai, MD, MSc

Department of Surgery, Ramathibodi Hospital, Mahidol University, Thailand

Abstract **The present article reviews the basic ideas behind summary or descriptive statistics. The types of data commonly seen in medical practice and research are introduced and described. Descriptive statistics appropriate for each type of data are presented and discussed. Motivations for these summaries are provided from a theoretical as well as practical point of view. The data set from a comparative study of laparoscopic methods for the treatment of morbid obesity is used as an illustration. Numerical exercises are provided at the end of the article. The introductory material in the article should help the reader begin his or her study of basic medical statistics.**

Key words: **descriptive statistics; summary statistics**

Introduction

The objective of the present article is to introduce the surgeon to the idea of descriptive or summary statistics. Some of the topics we will cover include: (1) the aims of summary statistics - why we need them; (2) types of numerical data; (3) types of summary statistics; and, (4) some motivations behind the choice of summary statistics. We also provide a few numerical exercises at the end of the article. Only rudimentary knowledge of statistics is assumed on the part of the reader. Although the article is self-contained, some familiarity with the first article in this series will be helpful.

Components of a measurement

The essential aim of applied statistics is to separate the "signal" from the "noise" in any given data set, to find "pattern" in apparent "randomness", that is, to separate the "good" from the "bad" and the "ugly" in the data.

Correspondence address : *Panuwat Lertsithichai, MD, MSc, Department of Surgery, Ramathibodi Hospital, Mahidol University, Thailand; Telephone: +66 2201 1315; Fax: +66 2201 1316; E-mail: raplt@mahidol.ac.th*

All measurements obtained in clinical practice, in a laboratory, or in clinical experiments have three components. Consider the measurement of height. The first component is the good one - the true height. The second component is the bias - the bad - which occurs through a variety of causes. The result of bias is a measurement of height which differs systematically from the true height. That is, the characteristic of bias is that it is repeatable and predictable given that it exists. For example, the meter stick used for measuring the height might actually be defective, being shorter than the standard meter. Then the measurement of height will be biased upwards. Every time we use this meter stick, it will always bias our results in a predictable manner. The ideal way to manage bias is to prevent its occurrence; otherwise, if we suspect that it exists, we must look for it, and get rid of it.

The third and last component is the ugly - random variation. However, beauty or ugliness "is in the eye of the beholder". Some might think randomness is beautiful. Nonetheless, random variation also results in measurements which differ from the true height. But the characteristic of random variation is such that each individual measurement will differ unpredictably (in both magnitude and direction) from the true value. The "causes" of random variation will often be unknown or will affect a measurement unpredictably. The usual way to manage random variation, or "noise", is to "control" or "allow" for it by means of appropriate research designs and statistical models. There is no getting rid of random variation.

By eliminating bias and controlling or allowing for random variation, we can use the height measurements to closely approximate the true height. In clinical research, the measurement of the difference between two or more treatments can be viewed in a similar manner as the measurement of height. If we use appropriate research methodology and statistical methods to eliminate bias and control random variation, we can demonstrate whether a treatment difference truly exists, and accurately measure its magnitude and direction.

Use of summary statistics

The use of appropriate summary statistics can help detect the true magnitude of a measure. For example, an appropriate summary can let us glimpse what the difference between two treatments might be, if such a difference exists. Also, by using summary statistics, patterns or relationships between numerical data are more easily discernable. This is something one might not see with raw data. As Sir Austin Bradford Hill puts it, "The publication ... of a long series of case results is not particularly helpful ... for it is impossible to detect, from the unsorted mass of raw material, relationships between the various factors at issue".[1]

A theoretical requirement for a set of summary statistics to be adequate is that it should contain all the necessary information within the data for statistical inference. That is, the set of summary statistics must be "sufficiently" representative of the data, in the statistical sense.[2] Sufficiency is defined relative to a specified statistical model for the data. For example, if each element of the data was assumed independent and identically distributed as a "Normal" variate, then the mean and variance (square of the standard deviation) would be sufficient statistics and hence adequate summary of the data set. When no fully specified statistical model is assumed, the choice of summary statistics is less clear.

Look at the raw data in Table 1, recorded from a study comparing two laparoscopic operations.[3] It is a part of a larger table comprising 65 patients. Of these, 31 patients underwent laparoscopic gastric banding and 34 underwent laparoscopic Roux-en-Y gastric bypass for the treatment of morbid obesity. The aim of the research study was to detect certain patterns or signals: i.e., that the two operations differ in terms of their effects on weight loss, and to look for risk factors associated with weight loss. Data collected included age, gender, type of operation, characteristics of operations, outcomes such as weight loss at various follow-up times, and a large quantity of other data not shown. It is obvious that the unprocessed, raw data in the table are of no use for detecting patterns of interest.

However, if we use the appropriate summary statistics, a pattern seems to emerge. Table 2 compares the summary statistics, which characterize the patients as well as their outcomes, between the two treatment groups. The last two rows summarizing the outcomes seem to suggest that, at one week, laparoscopic banding produced an average weight loss of 6.2 kg, while laparoscopic bypass produced an average weight loss of only 1.9 kg. So which treatment is better for morbidly fat people?

Table 1 A part of a data set comparing two laparoscopic operations (laparoscopic gastric banding and laparoscopic Roux-en-Y gastric baypass) for morbid obesity

Sex	Operative technique	Duration (min.)	Blood loss	ICU	LOS	wt. (kg.)	ht. (m.)	BMI	Ideal BW	Excess BW	1 wk.
M	Lap Banding	80	0	0	6	132	1.65	48.48485	71.25	60.75	124
M	Lap Banding	215	50	0	4	152	1.8	46.91358	80	72	143
M	Lap Banding	265	0	0	4	157	1.74	51.85626	75.5	81.5	150
F	Lap Banding	390	0	1	9	132	1.57	53.55187	62.5	69.5	
F	Lap Banding	145	10	0	3	103	1.58	41.25941	62.5	40.5	100
F	Lap Banding	225	80	1	5	140	1.68	49.60317	68.5	71.5	135
F	Lap Banding	140	50	0	2	115	1.59	45.48871	64	51	108
F	Lap Banding	190	0	0	6	90	1.45	42.80618	57	33	91
F	Lap Banding	165	50	0	3	131	1.68	46.4144	68.5	62.5	122
F	Lap Banding	120	0	0	3	102	1.6	39.84375	64	38	100
F	Lap Banding	85	0	0	4	97	1.51	42.54199	58.5	38.5	
F	Lap Banding	300	200	0	4	153	1.6	59.76563	64	89	144
F	Lap RYGB	165	0	0	5	116	1.6	45.3125	64	52	
F	Lap RYGB	350	200	0	5	140	1.52	60.59557	59.75	80.25	139
F	Lap Banding	125	0	0	3	102	1.55	42.45578	61	41	100
F	Lap RYGB	225	100	0	5	103	1.58	41.25941	62.5	40.5	103
M	Lap Banding	130	0	0	2	136	1.8	41.97531	80	56	127
M	Lap Banding	130	0	0	3	127	1.75	41.46939	77	50	121
F	Lap Banding	170	0	0	2	115	1.73	38.42427	71.5	43.5	
F	Lap Banding	190	0	0	2	149	1.72	50.36506	71.5	77.5	
M	Lap Banding	150	50	1	5	148	1.78	46.71127	78.5	69.5	142
M	Lap RYGB	290	50	1	8	163	1.71	55.74365	74	89	159

M = male; F = female; Lap = laparoscopic; RYGB = Roux-en-Y gastric bypass; LOS = length of stay; ICU=intensive care unit; wt = weight; ht = height; BMI = body mass index; BW = body weight; wk. = week

Table 2 Summary statistics of some variables from the full data set comparing two laparoscopic operations for morbid obesity. Observe the average weight loss at one week from the baseline

Characteristic	Lap banding (n = 31)	Lap RYGB (n = 34)	p-value
Age (years): mean (SD)	32.0 (11.3)	31.4 (9.8)	0.823
Sex (male): number (%)	15 (48)	13 (38)	0.409
Operative time (min): mean (SD)	164 (88.1)	301.6 (58.3)	< 0.001
Blood loss (ml): median (range)	0 (0 to 200)	100 (0 to 400)	< 0.001
ICU stay (yes): number (%)	6 (19)	23 (68)	< 0.001
Length of hospital stay (d): median (range)	3 (2 to 11)	6 (3 to 11)	< 0.001
Height (m): mean (SD)	1.69 (0.10)	1.68 (0.10)	0.571
Baseline body weight (kg): mean (SD)	128.4 (19.6)	134.1 (27.4)	0.340
Body weight at 1 week (kg): mean (SD)	122.2 (19.0)	132.2 (27.2)	0.159

SD = standard deviation; ICU = intensive care unit; Lap = laparoscopic; RYGB = Roux-en-Y gastric bypass; d = day. P-values were calculated using unpaired t-test, rank test, or chi-square test as appropriate.

There is an average difference in weight loss of 6.2 - 1.9 = 4.3 kg between the two groups, in favor of laparoscopic banding. If we remember our "good, bad and ugly" metaphor, this difference is a sum of three components. That is, there is a true difference, a bias component, and random variation. How can we be sure that 4.3 kg represents a true difference? If there is a large bias, or a significant random variation, or both, then 4.3 kg might represent either component or both, without there being any true difference whatsoever. Knowledge of study designs and their potential biases, as well as the theory of statistical methods, will help answer the question, but to a certain extent. This is not our present concern, however; we are raising some important issues here for future reference.

Types of numerical data in medical research

For practical purposes, we can classify numerical data or variables into three types.[4] The first is quantitative data. They may have a continuum of values, or may be intrinsically discrete, such as counts. They can be ranked; i.e., ordered. Examples include: age, temperature, blood pressure, and visual analogue scale (VAS) for pain. Take the variable age. Age can have a continuum of values - 10 years, 10.6 years, or 10.60008 years are all possible values for age. Age can be ranked; for example, 10 years is "older" than 7 years. Quantitative data can also have "interval" or "ratio" properties or both, but this will not be our concern.

The second type of data is nominal. They have only discrete "values" (categories) and cannot, by nature, be ranked, unless other considerations are taken into account. Examples include gender, types of operation and occupation.

The third type of data is ordinal. They have discrete values, but these values can be naturally ranked. They are, in a sense, "midway" between quantitative and nominal data. Examples include 4-level pain scale (severe pain, moderate pain, mild pain, no pain) or 5-level cosmetic scale (excellent, very good, good, poor, very poor or unacceptable). Consider the 4-level pain scale. The values are discrete because, for example, there are no possible values between "severe pain" and "moderate pain". Also, the values can be ranked. For example, "severe pain" is more painful than "moderate pain".

Summary statistics for quantitative data

What are the appropriate summary statistics for quantitative data? For quantitative data with "Normal" distributions, the mean and standard deviation are appropriate summaries. For quantitative data with non-Normal unimodal distributions, the median and range might be more appropriate.[5]

We assume that the student has some familiarity with statistical distributions. A distribution curve of any type of data can be visualized as a histogram: each point on the curve represents the relative frequency of observations ("Y" coordinate) which fall within the corresponding range of data values ("X" coordinate).[4]

Why should there be a distribution? Any collection of observations or measurements, for example, a set of height measurements or a collection of occupations in a group of people, has variability: not all observations are the same; not all persons have the same height or occupation. Without variability, there would be no distributions. The good, the bad and the ugly metaphor also apply in this situation. Variability in height in a group of people can be due to bias in measurement (the "bad"), but more importantly, variability is due to systematic factors (the "good"), and random variation (the "ugly"). Systematic factors related to height include, for example, age, socioeconomic status, gender and racial (genetic) makeup. These systematic factors are often the focus of research studies.

For quantitative variables with symmetric, bell-shaped distributions, the appropriate summary statistics include the mean (average) and the standard deviation (square-root of the variance). The motivation is that the bell-shaped unimodal distribution is consistent with a theoretical curve called the "Normal" or Gaussian distribution (Figure 1), whose characteristics can be sufficiently described by just two numbers: the mean and variance. Note that there must be at least two numbers to describe the distribution: one to describe the "central tendency", i.e., the mean (average), and one to describe the "spread", i.e., the variance or standard deviation.

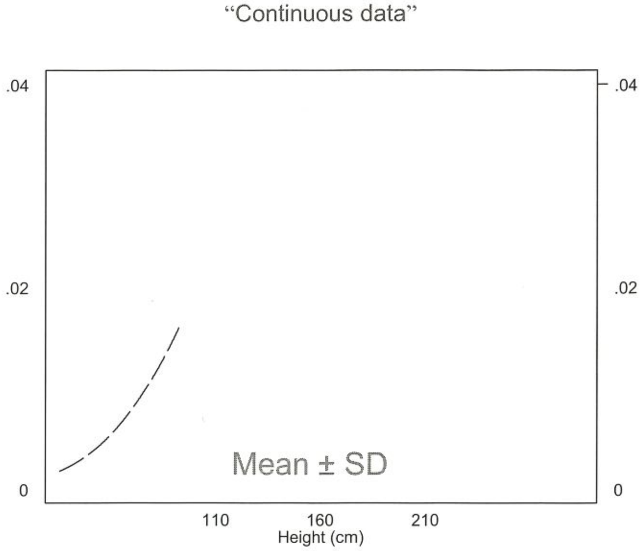

Figure 1 A relative frequency distribution of a quantitative and continuous Normal variate (a Normal or Gaussian distribution curve). The sufficient summary is the mean and standard deviation (SD).

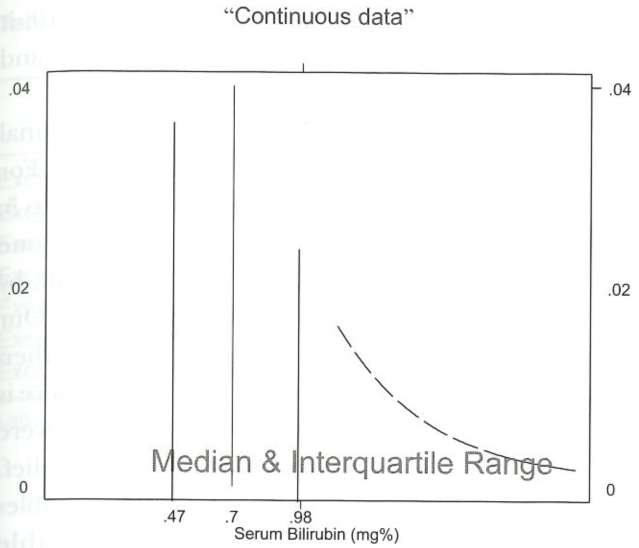

Figure 2 A relative frequency distribution of a quantitative and continuous non-Normal variate (a skewed unimodal distribution curve). An adequate summary is the median and range.

For non-symmetrical unimodal distributions, or "skewed" distributions (Figure 2), the mean and standard deviation are not appropriate summaries, or are not sufficient.[6,7] More often, the median, as a central tendency measure, and the range, as a measure of spread, are used.[5] The median is the value at the middle of a distribution, that is, when the data has been ranked, or ordered, from the smallest to the largest value. The overall range is simply the limits of the data: the smallest and largest values in the data.

One rationale or justification for using medians and various ranges for describing non-Normal data is that these summaries are based on ranks, or the ordering of the data. This is consistent with the use of statistical tests called non-parametric or "distribution-free" tests for non-Normal data, in which the ranks, instead of the actual values of the data, are sometimes used in constructing the tests. Thus, one justification for the use of medians and ranges as summaries is simply to be consistent with statistical inference procedures. The second rationale is that the value of the median, as a central tendency measure, is "robust" to outliers in the data, whereas the value of the mean is very sensitive to large outliers. This is especially relevant for skewed data and small-sized samples.

The third, theoretical rationale for using mediaus and range is that when the statistical model is not specified, no sufficient statistics are readily available except "order statistics".[8] Order statistics refers to the ranking of the data, and using this ranking to summarize the data. This theoretically justifies our use of medians and various ranges in most non-Normal situations.

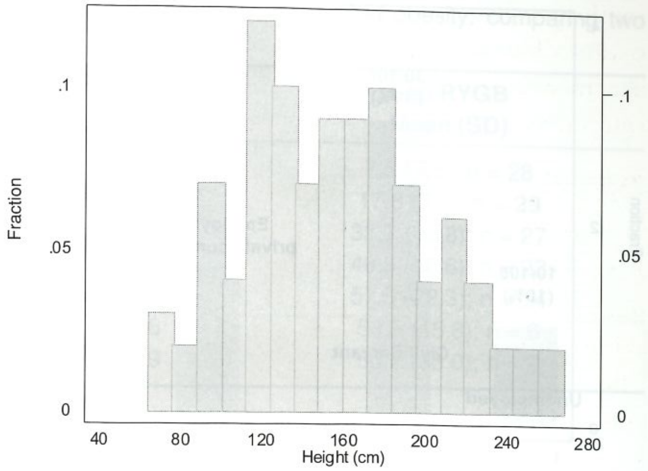

Figure 3 A histogram (discrete frequency distribution curve) of a simulated data set of 100 height observations, randomly drawn from a Normal distribution.

In actual practice, how do we determine, or how can we assume, that a quantitative variable has a Normal or non-Normal distribution? We recommend three simple determination rules.[5] With the appropriate statistical software, determine whether:

a. the mean is at least twice the standard deviation;

b. the mean is approximately the same or similar to the median;

c. the histogram is approximately "Normal-shaped".

Of the three rules the third is the most impractical, because the data are usually inadequate (small sample size, or large standard deviation) for constructing a histogram whose shape is either clearly Normal or clearly non-Normal. Therefore, only the first two rules are commonly used.

As an example, Figure 3 shows a histogram of a variate (or variable) randomly drawn 100 times from a Normal distribution, as simulated on a computer. The shape of the histogram is probably "Normal", but it is not entirely clear that it is so. Hence, the histogram is not very helpful in determining the type of data distribution in this case, even though the true underlying distribution is Normal.

Summaries for nominal data

For nominal data, the appropriate summaries are

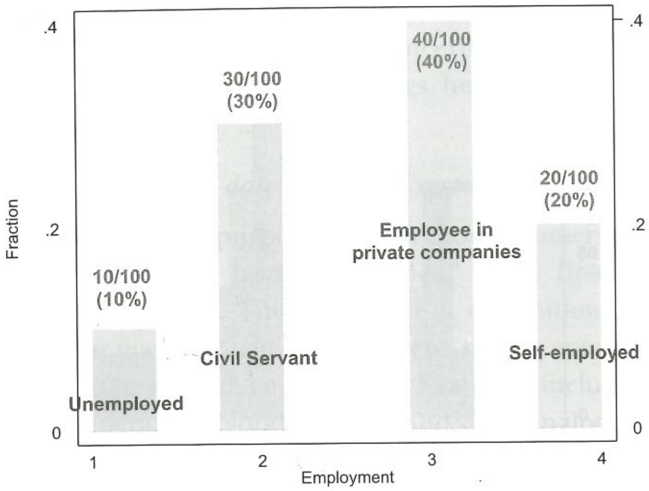

Figure 4 A histogram of the nominal categorical variate, employment. There are four categories with the corresponding summary statistics: counts and percentage (proportion) for each category.

counts and proportions. The category with the highest percentage or proportion of occurrence is the central tendency of the data. The spread of the data is demonstrated by the spread of the proportions across categories. Even if we assign numbers to the categories, it makes no sense to add them or to summarize in terms of means and standard deviations. There is no intrinsic way of reinterpreting a nominal variable to justify adding categories. This is in contrast to the ordinal variable, to be discussed below.

Theoretically, summarizing nominal data as counts and proportions also makes sense. If appropriate statistical models are assumed for the data, e.g. the Poisson or binomial model, then counts and proportions (i.e. totals) are sufficient statistics.[6]

Figure 4 presents a summary of the variable "occupation" in form of a histogram. Summary statistics are presented as well. The central tendency is apparently category 3, "employees in private companies". The spread of the data is across all four categories, with the smallest frequency in the "unemployed" category.

Summaries for ordinal data

For ordinal data, the choice of summary statistics might not be clear-cut. For a variable with only a few categories, e.g., no more than five categories, summarizing in terms of counts and proportions is probably the most informative. But when the number of categories is relatively large, say more than 10, it might be more intelligible to summarize as mean and standard deviation, or median and range.[5]

Some researchers worry that summarizing ordinal data as if they were quantitative is misleading.[5] For example, if a variable has 5 categories, labeled 1 to 5, does it make sense to add them? How does one interpret an average value of, say, 2.7? After all, by definition, no categories exist between 2 and 3. Our opinion is that if such summaries do mislead, then don't use them. But, in most cases, we believe there is no harm in summarizing as if the ordinal variable were quantitative. We provide three reasons for this belief.

The theoretical reason is that ordinal variables can have an underlying continuous variable interpretation.[9] Let us give an indirect example. Age is a continuous variable. Yet it can be categorized into 10 categories: category 1, age 0 to 10 years; category 2, 10 to 20 years, and so on, to category 10, > 90 years. The categorized age variable is then an ordinal variable, with categories 1 to 10. In this case, the categories can be added, and an average of 2.7 actually has meaning (i.e., the average age is 27 years), even if it is not a defined category! In a similar manner, any ordinal variable can have an underlying continuous variable interpretation such that the "average value" has some meaning.

The "consistency" reason, in simplified form, is that non-parametric tests are often used for testing ordinal variables.[10] Therefore, because these tests assume continuous distributions, it is "consistent practice" to summarize ordinal variables as if they were quantitative variables.

The practical reason, probably most pertinent for researchers, is that the mean, the median and the corresponding spread measures are best for visually detecting differences between groups. Because the aim of all research studies is to detect signals in the presence of noise, summaries that can help in detecting these signals should be used. For example, summarizing a 10-category ordinal variable using counts and proportions, and comparing two groups, will lead to the comparison of two columns of 10 numbers each. Only a few readers will have the patience to look for differences in two long columns of numbers. On the other hand, by summarizing the data in terms of two numbers per group (e.g., means and standard deviations), comparing between groups is much easier.

Table 3 Average excess body weight loss at various times after laparoscopic surgery for morbid obesity: comparing two operations

Outcomes	Lap banding Mean (SD)	Lap RYGB Mean (SD)
Excess body weight loss at 1 week (%)	11.3 (6.3); n = 21	7.5 (7.5); n = 28
Excess body weight loss at 1 month (%)	16.3 (8.5); n = 24	17.6 (7.5); n = 29
Excess body weight loss at 3 months (%)	21.4 (10.7); n = 23	32.7 (13.8); n = 27
Excess body weight loss at 6 months (%)	29.0 (18.5); n = 17	48.5 (16.6); n = 23
Excess body weight loss at 1 year (%)	36.8 (22.5); n = 15	57.5 (22.3); n = 14
Excess body weight loss at 1.5 years (%)	46.3 (18.8); n = 6	58.5 (15.8); n = 6
Excess body weight loss at 2 years (%)	47.8 (30.6); n = 9	50.7 (33.0); n = 2

Lap = laparoscopic; RYGB = Roux-en-Y gastric bypass; SD = standard deviation

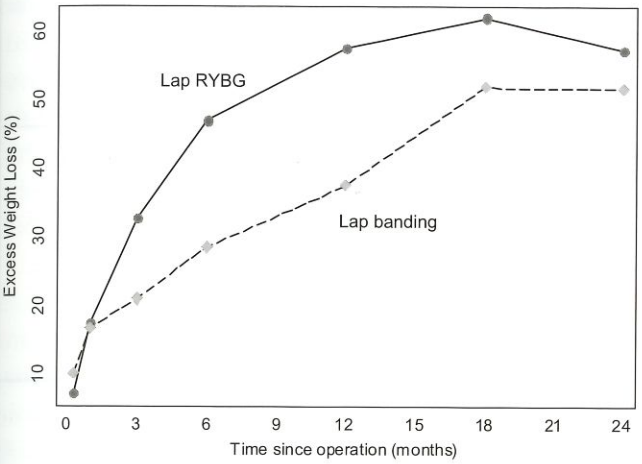

Figure 5 Comparing the serial excess weight loss between two laparoscopic operations for morbid obesity (laparoscopic (lap) gastric banding, dashed line; and laparoscopic R-en-Y gastric bypass (RYGB), solid line). The same data as presented in table 3.

Presenting statistical data and summaries

In presenting research data, summaries in terms of tables or in terms of graphics have their own advantages and disadvantages. In research articles, the most parsimonious way to present data is not to duplicate them. Indeed, all journals discourage duplication, unless the two presentations provide essential and complementary information. The choice between tabular and graphical presentations relies mainly on the criterion of clarity.

If, by looking at a table, the differences between groups are obvious, then the table is sufficient. However, when the contrast is between two or more treatments comparing observations at multiple time points, tables are a poor medium. In such cases, graphics are preferred, or should be included as well.

Table 3 presents a series of observations on weight loss at multiple time points, in a study comparing two operations for morbid obesity mentioned earlier. Visually, it is difficult to tell what is happening. In contrast, Figure 5 presents the same data in graphical form. Not only are the differences between the two operations clear-cut (or seemingly clear-cut) and informative, but the graph itself is nice to look at!

Summary of the article

We have presented a simplified version of how to summarize numerical data in medical research. We have classified numerical data commonly seen in medical research into three groups: quantitative, nominal, and ordinal. We have described the components of a measurement and the idea of statistical distributions. We provided concise guidelines and rules for summarizing numerical data in a variety of situations, as well as some theoretical and practical motivations for these rules. We hope that thereby students shall be better prepared to understand the basics of statistical methods.

REFERENCES

1. Hill AB. Principles of medical statistics. 8th ed. New York: Oxford University Press; 1966. p. 51.
2. Fisher RA. Theory of statistical estimation. Proc Camb Philol Soc 1925;22:700-25.
3. Vitoonpinyopab K, Angkoolpakdeekul T, Lertsithichai P. Short term outcomes of two laparoscopic procedures for morbid obesity (abstract). Rama Med J 2009; 32 (Suppl):421-2.
4. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell Scientific Publications; 1994. p. 14-26.

5. Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians; 1997.
6. Hogg RV, Craig AT. Introduction to mathematical statistics. 5th ed. New Jersey: Prentice-Hall; 1995. p. 307-40.
7. Pearson K. Contributions to the mathematical theory of evolution. Phil Trans R Soc Lond A 1894;185:71-110.
8. Casella G, Berger RL. Statistical inference. 2nd ed. Duxbury: Thomson Learning Inc.; 2002. p. 226-32, 275-6.
9. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman & Hall; 1989. p. 151-9.
10. Fisher LD, van Belle G. Biostatistics. New York: John Wiley & Sons; 1993. p. 317-8.

Exercise

Summarize the following data sets

1. 32, 34, 39, 42, 45, 47, 53, 58, 60, 60, 61, 63
2. 0,0,0,0,0,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,4,4
3. 1,1,2,2,2,2,3,3,3,4,4,4,4,4,4,5,5,5,6,6,7,7,7,9,9,9
4. 0,0,0,0,0,10,10,10,10,12,12,12,12,12,12,13,13
5. 68, 79, 80, 88, 98, 110, 134, 155, 160, 230, 347

Possible solutions

1. Probably quantitative data, compatible with a Normally distributed variate:
N = 12; Mean = 49.5; SD = 11.2

2. Probably nominal data. Nonetheless, summarize as the following is best:
N = 25
Category 0: number = 5 (20%)
Category 1: number = 9 (36%) [central tendency]
Category 2: number = 7 (28%)
Category 3: number = 2 (8%)
Category 4: number = 2 (8%)

3. Difficult to tell what type of data this might be. Suppose it is ordinal data. Then, with 8 categories, the following are probably appropriate:
N = 26
Mean = 4.5; SD = 2.4
Median = 4; range (1 to 9)

4. Almost impossible to say what type of data this is, until we know how they were derived or coded. But with a small number of categories, it is most informative to summarize as:
N = 17
Category 0: number = 5 (29%)
Category 10: number = 4 (24%)
Category 12: number = 6 (35%) [central tendency]
Category 13: number = 2 (12%)

5. Probably quantitative data, and obviously skewed. It is most appropriate to summarize as:
N = 11 ; median = 110 ; range (68 to 347)